PLOS | COMPUTATIONAL BIOLOGY

EDITORIAL

Ten Simple Rules for Selecting a Bio-ontology

James Malone[1]*, Robert Stevens[2], Simon Jupp[1], Tom Hancocks[1], Helen Parkinson[1], Cath Brooksbank[1]

1 European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom, **2** School of Computer Science, University of Manchester, Manchester, United Kingdom

* james@factbio.com

Citation: Malone J, Stevens R, Jupp S, Hancocks T, Parkinson H, Brooksbank C (2016) Ten Simple Rules for Selecting a Bio-ontology. PLoS Comput Biol 12(2): e1004743. doi:10.1371/journal.pcbi.1004743

Editor: Scott Markel, UNITED STATES

Published: February 11, 2016

Funding: EMBL and European Union grants Diachron (601043) and BioMedBridges (284209). National Institutes of Health NCBC grant for National Center for Biomedical Ontology [U54-HG004028]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Biologists and bioinformaticians now look to ontologies or software that uses ontologies as a means of standardising the way data are described, queried, and interpreted. Ontologies can be used for the annotation and curation of experimental datasets and, in data sharing, both within and beyond the confines of individual labs, organizations, and communities. Bio-ontologies are also commonly used in methods of analysis, particularly in gene set enrichment analysis [1], using ontologies such as the Gene Ontology. With modern high-throughput data-generation technologies, there is now, more than ever, a need to integrate data from these and other sources, and there is a concomitant need for ontologies—raising the question of how to choose a bio-ontology.

Over the past decade, a community has grown up around the success of efforts to harmonise the semantic description of biological entities, with ontologies exemplified in the emergence of the Open Biological and Biomedical Ontologies (OBO) Foundry [2]. These efforts were first led by the aforementioned Gene Ontology [3] and have expanded to ontologies that describe a significant range of the primary areas of biology and its science. Exploring bio-ontologies through browsers such as the Ontology Lookup Service [4] at the European Bioinformatics Institute and BioPortal [5] at the National Center for Biomedical Ontology (NCBO) —whose existence is itself a measure of the community size—shows there are over 400 ontologies containing, collectively, over 5 million classes (by classes, we mean ontological terms together with their associated descriptions and synonyms). These ontologies cover areas such as diseases [6], phenotypes [7], anatomy [8], experimental conditions [9,10], cell types [11], and bioinformatics software [12].

There are now many ontologies from which to choose, but which ontology should be chosen? In order to answer this question, we present ten simple rules that should help to guide the choice of a bio-ontology. The rules are designed to be useful for those wishing consume a bio-ontology. Users of bio-ontologies are varied in their profile and include data curators, application developers, and, of course, ontology developers who may be consuming part of an ontology in their own work.

Rule 1: The Ontology Should Be about a Specific Domain of Knowledge

Specifically, an ontology should provide coverage for the area it claims to describe. Although almost no ontology is complete, you should aim to find an ontology that describes a considerable amount of the area to which it lays claim. It should also describe the field of interest in such a way that extensions to cover missing areas are possible without a major rewriting of the ontology. Missing terms are to be expected, but if the ontology is missing large areas that are

key to describing the domain, then it may not be a suitable ontology. For instance, a disease ontology that does not include cancers would be inadequate if the diseases that you were aiming to describe included cancers. Furthermore, even if you personally don't have any cancer data to describe, you need to consider the notion that a disease ontology with such a large gap in it is unlikely to gain wide community adoption. Conversely, if an ontology claims to only describe a specific subset of a domain, or even just a local application, then it should do so appropriately and should not be considered an unsuitable ontology simply because it has a more limited scope; it may be less useful to the wider world, but this does not make it a bad ontology. One computational service that can help a user estimate whether or not an ontology can provide coverage is the NCBO Ontology Recommender [13]. Recommender measures whether an ontology from the NCBO BioPortal matches a given set of text based on a measure of coverage, which can help to inform whether a given ontology contains the terms a user might be expecting.

Rule 2: The Ontology Should Reflect Current Understanding of Biological Systems

Unless the aim of the ontology is to capture a historic viewpoint (a legitimate objective), then it should reflect current science, or at least not contradict it. For instance, the old dogma of DNA to RNA without feedback would no longer be accepted in modern biology. It is better to make statements that are too broad but remain correct rather than make specific statements that are wrong. The correctness of an ontology is often evaluated using techniques such as competency questions. [14]. Competency questions are queries that are required to be answerable by an ontology, with the returned answers thus demonstrating whether or not an ontology is giving correct, expected answers. In the simplest form, a user may ask for subtypes of a class, e.g., "What are the subclasses of fat cell?", but this can also be more complex depending upon needs, e.g., "Which cell lines are derived from human, epithelial cells and are taken from melanoma samples?" Correct answers suggest the ontology is reflecting current science correctly.

Rule 3: The Ontology Classes and Relationships Should Persist

One of the primary use cases of an ontology is to describe biomedical data through annotations; disconnecting the descriptions of this data from their semantics through the deletion of the ontology identifiers undermines the advantage of using ontologies in the first instance. This is crucial if these ontologies' annotations are being used for data sharing, integration, or analysis. Identifiers in most biomedical ontologies are formed using accessioned IDs rather than textual labels, with the intent of removing potential ID clashes and decoupling the textual part of a class (i.e., the label) from the identifier referring to it. This has the advantage of enabling small modifications to a class label without affecting the class identifier, where the class is still referring to the same entity. In cases in which the identifier is a Uniform Resource Identifier (URI), these should resolve to provide both human-readable and machine-interpretable information. Services like Identifiers.org provide a URI resolution service for many biomedical ontologies. Identifiers should be maintained, and if it is necessary to remove a class, it should be labelled as "obsolete" rather than simply deleted. Maintaining this audit trail is the sign of a well-managed ontology—deleting identifiers is the sign of a poorly managed ontology.

Rule 4: Classes Should Contain Textual Definitions

This is crucial for users who come to an ontology trying to understand what a particular class is describing. It may, on occasion, be obvious—"*Homo sapiens*," for instance. On others, it is critical—a cell line named "Bas666" could be difficult to interpret. An additional sign of a

suitable ontology is that it contains appropriate synonyms (e.g., "human" for "*Homo sapiens*") and related alternative terms (e.g., in Gene Ontology, use of narrow synonyms such as "type I programmed cell death" for the label "apoptotic process"), since language can differ between communities and specialities even though the underlying class being described is the same thing. As well as textual definitions, many ontologies also contain logical descriptions of the class that are amenable to computational interpretation. These descriptions use rules, or "axioms," to relate a class to other classes, such as describing that a heart is part of the cardiovascular system. Whether or not such computational aspects are necessary for a particular use case should form part of the decision when selecting a bio-ontology.

Rule 5: Textual Definitions Should Be Written for Domain Experts

Creators of ontologies often fall into the trap of defining classes using ontology jargon (often philosophical in nature). This may make them understandable to ontologists and/or philosophers, but this is not useful if the language used means nothing to the ontology's user community. A good ontology will reflect commonly used nomenclature in naming classes within it. Similarly, textual definitions should also reflect common language used in the biological domain. Textual definitions and labels that include ontology jargon are the sign of an unsuitable ontology. Ontologists should accurately describe the biomedical domain without modifying it.

Rule 6: The Ontology Should Be Developed by the Community but Not Incapacitated by It

Reflecting current science is a difficult task, given the growing knowledge of the breadth and depth of entities in the domain. Gaining community consensus is a noble cause and should help to reflect current science correctly and enhance opportunities for wide adoption of an ontology. It is also almost always better to work towards getting entities added into an existing ontology that is supported by a community rather than inventing a new ontology. Engaging with the community, however, should not deflect from the task of developing an ontology. Decision making—should a user ask for a new class, for example—should not take months while consensus is obtained. Similarly, a lone gatekeeper making all the decisions about what happens within an ontology is also a bad sign. Most ontologies will have a public forum for dealing with user requests, and looking at mailing list archives or issue trackers (e.g., Gene Ontology http://geneontology.org/page/go-mailing-lists) will provide insight on how the ontology is being developed.

Another aspect of collaborating is that of compromise. Typically, everyone has an opinion on the science in which they are interested, and typically they don't all align, so there is an element of compromise to selecting an ontology. A favoured label might not exist in the ontology, but rejecting wholesale a community-developed ontology in favour of inventing a de novo artefact with one's own favourite labels is not always the best option. As above though, there are circumstances under which this may be ultimately the better option; here, the balance is in weighing up requirements and making a judgement based on what is most important. Wider integration with a community is a good thing when it works.

Rule 7: The Ontology Should Be under Active Development

An ontology should have a dedicated presence on the web, such as a project website that provides information on how to contact the developers and contribute to the ontology. Any associated mailing lists or version control systems can be used to gauge recent activity on the ontology. Recent work [15] has shown that it is possible to describe how actively an ontology is

maintained and in what way it is being modified. In general, an ontology that is not actively developed and has not been updated for many months or years is unlikely to respond to new requirements should they occur.

Rule 8: Previous Versions Should Be Available

Given the changing nature of data and, hopefully, of the ontology as it updates to reflect current knowledge, data annotations can become out of date. An ontology should provide a clear versioning and release policy. It is important to be able to access older versions of an ontology so the context of data descriptions can be understood. This also relates to Rule 3 about not deleting ontology classes, and in turn both rules relate to enabling reproducibility of data analysis. Being unable to trace provenance of data annotations made with an older version of an ontology is a barrier to future reproducibility; selecting an ontology that maintains previous versions is therefore an important consideration.

Rule 9: Open Data Requires Open Ontologies

An important consideration is whether or not the ontology is being selected for use with open data with the intention of wider sharing. Using an ontology that is restrictive in licensing can also have an impact on the data described with such an ontology, restricting access to the semantics which are necessary to understand it. If data sharing is the aim, then using ontologies with permissive licenses should be a priority. Permissive licenses, such as those developed by Creative Commons (https://creativecommons.org/licenses/), can be used to communicate both how the work (ontology) can be exploited and how attribution should be given. One of the important outcomes of the standardization efforts from the OBO Foundry has been the widespread use of the OBO Relation Ontology (RO), an open ontology of biomedical relationships. The widespread use of RO has led to a de facto standard in much of the bio-ontology world, which has had positive implications on integration of resources, facilitated by the open license with which the RO is released.

Rule 10: Sometimes an Ontology Is Not Needed at All

Ontologies provide a means of "knowing" what is being described in a data set. There is, however, more than one way to capture such knowledge. Before embarking on using or indeed making a bio-ontology, you need to decide whether an ontology is really what is needed. In the broadest terms, we are talking about knowledge organisation systems of which there are numerous types of useful resources: glossaries, taxonomies, thesauri, ontologies, and terminologies. As a growing discipline, there is a temptation to suggest that using biomedical ontologies will offer some advantage. Ontologies offer advantages over other knowledge systems—they enable both computational use and human understanding, they can contain multiple classification axes of classes as well as formal descriptions of how classes relate to one another, and can include rich vocabularies of labels, synonyms, and textual definitions. If these are desirable selection criteria, then an ontology should be considered. Ontologies do also come with computational overheads, however, and can be complex to understand. Languages such as the Web Ontology Language (OWL) [16] utilise description logics, which are technically challenging. Other resources such as a vocabulary do not offer the sorts of classification and rich computational descriptions of an ontology but are often much simpler to understand. Let your requirements guide you; ontologies are not a panacea—sometimes one isn't needed at all.

Conclusion

Bio-ontologies represent an important tool for describing metadata, an increasingly important consideration as the scientific community aims for open, reusable data. It is perhaps no surprise then that in the Ten Simple Rules for the Care and Feeding of Scientific Data [17], the word "metadata" appeared 11 times. The choice of which ontology to pick and even when to use one is not always straightforward, as demonstrated by the number of times the authors are asked to recommend a particular ontology for a given problem. The single most important consideration in selecting a bio-ontology is to understand requirements first before deciding to engage with a particular ontology or indeed before minting one's own ontology. By identifying needs and selecting current ontologies using the above rules, it is possible to reach a conclusion as to whether or not a resource is useful to a given user. Moreover, reusing ontologies that similarly satisfy another user's needs helps to spread the burden of development across the community and ensure we don't end up with islands of metadata, undermining the efforts of openness and sharing.

Acknowledgments

We would like to thank all those involved in building open access bio-ontologies, especially those in the OBO community, and Judith Blake and Mark Musen for their comments on improving this article.

References

1. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009; 37(1):1–13. http://nar.oxfordjournals.org/content/37/1/1.abstract. doi: 10.1093/nar/gkn923 PMID: 19033363
2. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007 November; 25(11):1251–1255. doi: 10.1038/nbt1346 PMID: 17989687
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25–29. PMID: 10802651
4. Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H. The Ontology Lookup Service: bigger and better. Nucleic Acids Research. 2010; 38(suppl 2):W155–W160. http://nar.oxfordjournals.org/content/38/suppl_2/W155.abstract.
5. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research. 2009; 37(suppl 2):W170–W173.
6. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research. 2012; 40(D1):D940–D946. http://nar.oxfordjournals.org/content/40/D1/D940.abstract.
7. Mabee PM, Ashburner M, Cronk Q, Gkoutos GV, Haendel M, Segerdell E, et al. Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology and Evolution. 2007; 22(7):345–350. http://www.sciencedirect.com/science/article/pii/S0169534707001048. PMID: 17416439
8. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA, et al. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012; 13(1):R5. doi: 10.1186/gb-2012-13-1-r5 PMID: 22293552
9. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010; 26(8):1112–1118. http://bioinformatics.oxfordjournals.org/content/26/8/1112.abstract. doi: 10.1093/bioinformatics/btq099 PMID: 20200009
10. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010; 1(Suppl 1):S7. doi: 10.1186/2041-1480-1-S1-S7 PMID: 20626927
11. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome biology. 2005; 6(2):R21. PMID: 15693950

12. Malone J, Brown A, Lister AL, Ison J, Hull D, Parkinson H, et al. The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. Journal of Biomedical Semantics. 2014; 5(25).
13. Jonquet C, Musen M, Shah N. Building a biomedical ontology recommender web service. Journal of Biomedical Semantics. 2010; 1(S1).
14. Azzaoui K, Jacoby E, Senger S, Rodríguez EC, Loza M, Zdrazil B, et al. Scientific competency questions as the basis for semantically enriched open pharmacological space development. Drug Discovery Today. 2013; 18(17–18):843–852. doi: 10.1016/j.drudis.2013.05.008 PMID: 23702085
15. Malone J, Stevens R. Measuring the level of activity in community built bio-ontologies. Journal of Biomedical Informatics. 2013; 46(1):5–14. http://www.sciencedirect.com/science/article/pii/S153204641200055X. doi: 10.1016/j.jbi.2012.04.002 PMID: 22554701
16. OWL Working Group W. OWL 2 Web Ontology Language: Document Overview. W3C; 11 December 2012. http://www.w3.org/TR/owl2-overview/.
17. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol. 2014; 10(4):e1003542. doi: 10.1371/journal.pcbi.1003542 PMID: 24763340